Amendments to the Specification:

Please substitute the following title for the title on line 12 of page 1.

METHODS OF IDENTIFYING PROTEASE MODULATORS

Please substitute the following paragraph for the paragraph bridging pages 1-2.

"Protease," "proteinase," and "peptidase" are synonymous terms applying to all enzymes that hydrolyze peptide bonds, *i.e.* proteolytic enzymes. Proteases are an exceptionally important group of enzymes in medical research and biotechnology. They are necessary for the survival of all living creatures, and are encoded by 1-2% of all mammalian genes. Rawlings and Barrett (MEROPS: the peptidase database. *Nucleic Acids Res.*, 1999, 27:325-331) ~~(http://www.babraham.co.uk/Merops/ Merops.htm~~ (Which is incorporated herein by reference in its entirety including any figures, tables, or drawings.) have classified peptidases into 157 families based on structural similarity at the catalytic core sequence. These families are further classed into 26 clans, based on indications of common evolutionary relationship. Peptidases play key roles in both the normal physiology and disease-related pathways in mammalian cells. Examples include the modulation of apoptosis (caspases), control of blood pressure (renin, angiotensin-converting enzymes), tissue remodeling and tumor invasion (collagenase), the development of Alzheimer's Disease (β-secretase), protein turnover and cell-cycle regulation (proteosome), and inflammation (TNF-α convertase). (Barrett *et al.*, Handbook of Proteolytic Enzymes, 1998, Academic Press, San Diego which is incorporated herein by reference in its entirety including any figures, tables, or drawings.)

Please substitute the following paragraph for the first full paragraph on page 2.

Peptidases are classed as either exopeptidases or endopeptidases. The exopeptidases act only near the ends of polypeptide chains: aminopeptidases act at the free N-terminus and carboxypeptidases at the free C-terminus. The endopeptidases are divided, on the basis of

-2-

their mechanism of action, into six sub-subclasses: aspartyl endopeptidases (3.4.23), cysteine endopeptidases (3.4.22), metalloendopeptidases (3.4.24), serine endopeptidases (3.4.21), threonine endopeptidases (3.4.25), and a final group that could not be assigned to any of the above classes (3.4.99). (Enzyme nomenclature and numbering are based on "Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) 1992, (http://www.chem.qmw.ac.uk/iubmb/enzyme/EC34/intro.html).)

Please substitute the following paragraph for the paragraph bridging pages 5-6.

A first aspect of the invention features an identified, isolated, enriched, or purified nucleic acid polypeptide molecule having an amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, SEQ ID NO:70, SEQ ID NO:71, SEQ ID NO:72, SEQ ID NO:73, SEQ ID NO:74, SEQ ID NO:75, SEQ ID NO:76, SEQ ID NO:77, SEQ ID NO:78, SEQ ID NO:79, SEQ ID NO:80, SEQ ID NO:81, SEQ ID NO:82, SEQ ID NO:83, SEQ ID NO:84, SEQ ID NO:85, SEQ ID NO:86, SEQ ID NO:87, SEQ ID NO:88, SEQ ID NO:89, SEQ ID NO:90, SEQ ID NO:91, SEQ ID NO:92, SEQ ID NO:93, SEQ ID NO:94, SEQ ID NO:95, SEQ ID NO:96, SEQ ID NO:97, SEQ ID NO:98, SEQ ID NO:99, SEQ ID NO:100, SEQ ID NO:101, SEQ ID NO:102, SEQ ID NO:103, SEQ ID NO:104, SEQ ID NO:105, SEQ ID NO:106, SEQ ID NO:107, SEQ ID NO:108, SEQ ID NO:109, SEQ ID NO:110, SEQ ID NO:111, SEQ ID NO:112, SEQ ID NO:113, SEQ ID NO:114, SEQ ID NO:115, SEQ ID NO:116, SEQ ID NO:117 and SEQ ID NO:118 and biological domains thereof.

Please substitute the following paragraph for the paragraph bridging pages 76-77.

A partial list of proteases known to belong to this large and important family include: blood coagulation factors VII, IX, X, XI and XII; thrombin; plasminogen; complement

components C1r, C1s, C2; complement factors B, D and I; complement-activating component of RA-reactive factor; elastases 1, 2, 3A, 3B (protease E); hepatocyte growth factor activator; glandular (tissue) kallikreins including EGF-binding protein types A, B, and C; NGF-γ chain, γ-renin, and prostate specific antigen (PSA); plasma kallikrein; mast cell proteases; myeloblastin (proteinase 3) (Wegener's autoantigen); plasminogen activators (urokinase-type, and tissue-type); and the trypsins I, II, III, and IV. These peptidases play key roles in coagulation, tumorigenesis, control of blood pressure, release of growth factors, and other roles. (http://www.babraham.co.uk/Merops/Merops.htm).

Please substitute the following paragraph for the paragraph on page 123.

Table 2 lists the following features of the genes described in this patent application: chromosomal localization, single nucleotide polymorphisms (SNPs), representation in dbEST, and repeat regions. From left to right the data presented is as follows: "Gene Name", "ID#na", "ID#aa", "FL/Cat", "Superfamily", "Group", "Family", "Chromosome", "SNPs", "dbEST_hits", & "Repeats". The contents of the first 7 columns (i.e.,. "Gene Name", "ID#na", "ID#aa", "FL/Cat", "Superfamily", "Group", "Family") are as described above for Table 1. "Chromosome" refers to the cytogenetic localization of the gene. Information in the "SNPs" column describes the nucleic acid position and degenerate nature of candidate single nucleotide polymorphisms (SNPs; please see table of polymorphism below). These SNPs were identified by blastn of the DNA sequence against the database of single nucleotide polymorphisms maintained at NCBI (http://www.ncbi.nlm.nih.gov/SNP/snpblastByChr.html). "dbEST hits" lists accession numbers of entries in the public database of ESTs (dbEST, http://www.ncbi.nlm.nih.gov/dbEST/index.html) that contain at least 150 bp of 100% identity to the corresponding gene. These ESTs were identified by blastn of dbEST. "Repeats" contains information about the location of short sequences, approximately 20 bp in length, that are of low complexity and that are present in several distinct genes.

-4-

Please substitute the following paragraph for the paragraph on page 127.

Table 4 describes the results of Smith Waterman similarity searches (Matrix: Pam100; gap open/extension penalties 12/2) of the amino acid sequences against the NCBI database of non-redundant protein sequences (http://www.ncbi.nlm.nih.gov/Entrez/protein.html). The column headings are: "Gene Name", "ID#na", "ID#aa", "FL/Cat", "Superfamily", "Group", "Family", "Pscore", "aa_length", "aa_ID_match", "%Identity", "%Similar", "ACC#_nraa_match", and "Description". The contents of the first 7 columns (i.e.,. "Gene Name", "ID#na", "ID#aa", "FL/Cat", "Superfamily", "Group", "Family") are as described above for Table 1. "Pscore" refers to the Smith Waterman probability score. This number approximates the chance that the alignment occurred by chance. Thus, a very low number, such as 2.10E-64, indicates that there is a very significant match between the query and the database target. "aa_length" refers to the length of the protein in amino acids. "aa_ID_match" indicates the number of amino acids that were identical in the alignment. "% Identity" lists the percent of amino acids that were identical over the aligned region. "% Similarity" lists the percent of amino acids that were similar over the alignment. "ACC#nraa_match" lists the accession number of the most similar protein in the NCBI / database of non-redundant proteins. "Description" contains the name of the most similar protein in the NCBI database of non-redundant proteins.

Please substitute the following paragraph for the last paragraph on page 129.

Novel proteases were identified from the Celera human genomic sequence databases, and from the public Human Genome Sequencing project (http://www.ncbi.nlm.nih.gov/) using hidden Markov models (HMMR). The genomic database entries were translated in six open reading frames and searched against the model using a Timelogic Decypher box with a Field programmable array (FPGA) accelerated version of HMMR2.1. The DNA sequences encoding the predicted protein sequences aligning to the HMMR profile were extracted from the original genomic database. The nucleic acid sequences were then clustered using the Pangea Clustering tool to eliminate repetitive entries. The putative protease sequences were

then sequentially run through a series of queries and filters to identify novel protease sequences. Specifically, the HMMR identified sequences were searched using BLASTN and BLASTX against a nucleotide and amino acid repository containing known human proteases and all subsequent new protease sequences as they are identified. The output was parsed into a spreadsheet to facilitate elimination of known genes by manual inspection. Two models were used, a "complete" model and a "partial" or Smith Waterman model. The partial model was used to identify sub-catalytic domains, whereas the complete model was used to identify complete catalytic domains. The selected hits were then queried using BLASTN against the public NRNA and EST databases to confirm they are indeed unique.

Please substitute the following paragraph for the first paragraph on page 130.

Extension of partial DNA sequences to encompass the longer sequences, including full-length open-reading frame, was carried out by several methods. Iterative blastn searching of the cDNA databases listed in Table 5 was used to find cDNAs that extended the genomic sequences. "LifeGold" databases are from Incyte Genomics, Inc (http://www.incyte.com/). NCBI databases are from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). All blastn searches were conducted using a penalty for a nucleotide mismatch of −3 and reward for a nucleotide match of 1. The gapped blast algorithm is described in: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402).

Please substitute the following paragraph for the last paragraph on page 131.

Another method involved using the Genewise program (http://www.sanger.ac.uk/Software/Wise2/) to predict potential ORFs based on homology to the closest orthologue/homologue. Genewise requires two inputs, the homologous protein,

and genomic DNA containing the gene of interest. The genomic DNA was identified by blastn searches of Celera and Human Genome Project databases. The orthologs were identified by blastp searches of the NCBI non-redundant protein database (NRAA). Genewise compares the protein sequence to a genomic DNA sequence, allowing for introns and frameshifting errors.

Please substitute the following paragraph for the paragraph bridging pages 132-133.

The sources for the sequence information used to extend the genes in the provisional patents are listed below. For genes that were extended using Genewise, the accession numbers of the protein ortholog and the genomic DNA are given. (Genewise uses the ortholog to assemble the coding sequence of the target gene from the genomic sequence). The amino acid sequences for the orthologs were obtained from the NCBI non-redundant database of proteins. (http://www.ncbi.nlm.nih.gov/Entrez/protein.html). The genomic DNA came from two sources: Celera and NCBI-NRNA, as indicated below. cDNA sources are also listed below. All of the genomic sequences were used as input for Genscan predictions to predict splice sites [Burge and Karlin, JMB (1997) 268(1):78-94)]. Abbreviations: HGP: Human Genome Project; NCBI, National Center for Biotechnology Information.

Please substitute the following paragraph for the last paragraph on page 170.

SGPr480, SEQ ID NO:14, SEQ ID NO:73 encodes a protein that is 1604 amino acids long. It is classified as a Cysteine protease, of the UCH2b family. The protease domain(s) in this protein match the hidden Markov profile for a Ubiquitin carboxyl-terminal hydrolase family 2b (PF00443), from amino acid 1506 to amino acid 1566. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 72. Other domains identified within this protein are: UCH2b (PF00442) from 734 to 765; and two EF hands (PF00036) from 232 to 260, and from 268 to 296. Many calcium-binding

proteins belong to the same evolutionary family and share a type of calcium-binding domain known as the EF-hand (see http://www.expasy.ch/cgi-bin/prosite-search-ac?PDOC00018). This type of domain consists of a twelve residue loop flanked on both side by a twelve residue alpha-helical domain. In an EF-hand loop the calcium ion is coordinated in a pentagonal bipyramidal configuration. This protein has a putative CAAX motif (portion of SEQ ID NO: 73) (CVLQ) which may direct it to the membrane fraction. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 0; number of identical amino acids = 1272; percent identity = 99%; percent similarity = 99%; the accession number of the most similar entry in NRAA is NP_115971.1; the name or description, and species, of the most similar protein in NRAA is: ubiquitin specific protease [Homo sapiens].

Please substitute the following paragraph for the paragraph bridging pages 173-174.

SGPr359, SEQ ID NO:20, SEQ ID NO:79 encodes a protein that is 483 amino acids long. It is classified as a Metalloprotease protease, of the PepM10 family. The protease domain(s) in this protein match the hidden Markov profile for a Peptidase_M10 (PF00413), from amino acid 44 to amino acid 212. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 168. Other domains identified within this protein are: 3 x Hemopexin (PF00045) domains from 302 to 403. Hemopexin is a serum glycoprotein that binds heme and transports it to the liver for breakdown and iron recovery, after which the free hemopexin returns to the circulation. Hemopexin-like domains have been found in two types of proteins: - in vitronectin, a cell adhesion and spreading factor found in plasma and tissues and in most members of the matrix metalloproteinases family (matrixins), including MMP-1, MMP-2, MMP-3, MMP-8, MMP-9, MMP-10, MMP-11, MMP-12, MMP-13, MMP-14, MMP-15, MMP-16, MMP-17, MMP-18, MMP-19, MMP-20, MMP-24, and MMP-25 (see http://www.expasy.ch/cgi-bin/prosite-search-ac?PDOC00023). The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with

this protein sequence yielded the following results: Pscore = 0; number of identical amino acids = 483; percent identity = 100%; percent similarity = 100%; the accession number of the most similar entry in NRAA is NP_004762.1; the name or description, and species, of the most similar protein in NRAA is: matrix metalloproteinase 20 preproprotein; enamelysin [Homo sapiens]. This protein has a transmembrane domain from amino acid 7 to amino acid 29. This may function as a signal peptide.

Please substitute the following paragraph for the paragraph bridging pages 185-186.

SGPr559, SEQ ID NO:44, SEQ ID NO:103 encodes a protein that is 454 amino acids long. It is classified as a Serine protease, of the trypsin family. The protease domain(s) in this protein match the hidden Markov profile for a trypsin (PF00089), from amino acid 217 to amino acid 444. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 259. Other domains identified within this protein are: Low-density lipoprotein receptor domain class A (PF00057), from amino acid 71 to 109. LDL-receptors the class A domains form the binding site for LDL and calcium. The acidic residues between the fourth and sixth cysteines are important for high-affinity binding of positively charged sequences in LDLR's ligands. The repeat has been shown to consist of a beta-hairpin structure followed by a series of beta turns (see http://www.expasy.ch/cgi-bin/get-prodoc-entry?PDOC00929). The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 1.40E-288; number of identical amino acids = 454; percent identity = 100%; percent similarity = 100%; the accession number of the most similar entry in NRAA is NP_076927.1; the name or description, and species, of the most similar protein in NRAA is: transmembrane protease, serine 3 [Homo sapiens]. This protein has a transmembrane domain from amino acid 49 to amino acid 71.

Please substitute the following paragraph for the first full paragraph on page 188.

SGPr524_1, SEQ ID NO:49, SEQ ID NO:108 encodes a protein that is 850 amino acids long. It is classified as a Serine protease, of the trypsin family. The protease domain(s) in this protein match the hidden Markov profile for a trypsin (PF00089), from amino acid 613 to amino acid 842. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 259. Other domains identified within this protein are: three Low-density lipoprotein receptor domain class A domains (PF00057) from 489 to 603. LDL-receptors the class A domains form the binding site for LDL and calcium. The acidic residues between the fourth and sixth cysteines are important for high-affinity binding of positively charged sequences in LDLR's ligands. The repeat has been shown to consist of a beta-hairpin structure followed by a series of beta turns ~~(see http://www.expasy.ch/cgi-bin/get-prodoc-entry?PDOC00929)~~. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 1.30E-79; number of identical amino acids = 193; percent identity = 41%; percent similarity = 55%; the accession number of the most similar entry in NRAA is BAB23684.1; the name or description, and species, of the most similar protein in NRAA is: (AK004939) putative [Mus musculus]. This protein has a transmembrane domain from amino acid 77 to amino acid 99.

Please substitute the following paragraph for the paragraph bridging pages 190-191.

SGPr551, SEQ ID NO:54, SEQ ID NO:113 encodes a protein that is 802 amino acids long. It is classified as a Serine protease, of the trypsin family. The protease domain(s) in this protein match the hidden Markov profile for a trypsin (PF00089), from amino acid 568 to amino acid 797. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 259. Other domains identified within this protein are: three low-density lipoprotein receptor domain class A domains (PF00057) from 447 to 559. LDL-receptors the class A domains form the binding site for LDL and calcium. The

acidic residues between the fourth and sixth cysteines are important for high-affinity binding of positively charged sequences in LDLR's ligands. The repeat has been shown to consist of a beta-hairpin structure followed by a series of beta turns (see http://www.expasy.ch/cgi-bin/get-prodoc-entry?PDOC00929). The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 0; number of identical amino acids = 675; percent identity = 84%; percent similarity = 90%; the accession number of the most similar entry in NRAA is BAB23684.1; the name or description, and species, of the most similar protein in NRAA is: (AK004939) putative [Mus musculus]. This protein has a transmembrane domain from amino acid 44 to amino acid 66. This region could function as a signal peptide.

Please substitute the following paragraph for the paragraph bridging pages 203-204.

Several sources were used to find information about the chromosomal localization of each of the genes described in this patent. First, the Celera Browser was used to map the genes. Alternatively, the accession number of a genomic contig (identified by BLAST against NRNA) was used to query the Entrez Genome Browser (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/MapViewerHelp.html), and the cytogenetic localization was read from the NCBI data. References for association of the mapped sites with chromosomal amplifications found in human cancer can be found in: Knuutila, et al., Am J Pathol, 1998, 152:1107-1123. Information on mapped positions was also obtained by searching published literature (at NCBI, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) for documented association of the mapped position with human disease.

Please substitute the following paragraph for the paragraph at lines 11-23 on page 211.

The most common variations in human DNA are single nucleotide polymorphisms (SNPs), which occur approximately once every 100 to 300 bases. Because SNPs are expected to facilitate large-scale association genetics studies, there has recently been great interest in SNP discovery and detection. Candidate SNPs for the genes in this patent were identified by blastn searching the nucleic acid sequences against the public database of sequences containing documented SNPs (dbSNP, at NCBI, ~~http://www.ncbi.nlm.nih.gov/SNP/snpblastpretty.html~~). dbSNP accession numbers for the SNP-containing sequences are given. SNPs were also identified by comparing several databases of expressed genes (dbEST, NRNA) and genomic sequence (i.e., NRNA) for single basepair mismatches. The results are shown in Table 1, in the column labeled "SNPs". These are candidate SNPs – their actual frequency in the human population was not determined. The code below is standard for representing DNA sequence: